PARAMETERS THAT DEPEND LINEARLY ON A CONDITIONAL MEAN

There are many interesting parameters that depend linearly on a regression, i.e. parameters that are a linear function of a conditional mean.

One of these is Example 2, the average treatment effect (ATE).

Let $D \in \{0, 1\}$ be a binary treatment variable and Z a vector of covariates.

Suppose that there is an outcome variable Y satisfying

$$Y = DY(1) + (1 - D)Y(0),$$

where $Y(1)$ and $Y(0)$ are unobserved variables that are the potential outcomes with and without treatment respectively.

Suppose also that $Y(1)$ and $Y(0)$ are independent of D conditional on Z and the propensity score $\pi_0(Z) = \mathsf{Pr}(D = 1|Z)$ satisfies $0 < \pi_0(Z) < 1$ with probability 1.

It is well known that for $\gamma_0(D, Z) = E[Y|D, Z]$ the ATE is

$$\theta_0 := E[Y(1) - Y(0)] = E[\gamma_0(1, Z) - \gamma_0(0, Z)].$$

Example 2: The ATE is

$$\theta_0 = E[\gamma_0(1, Z) - \gamma_0(0, Z)],\ \gamma_0(D, Z) = E[Y|D.Z].$$

This parameter is the expectation of the linear function $\gamma_0(1, Z) - \gamma_0(0, Z)$ of the conditional mean.

This also can be interpreted as the effect of D on $E[Y|D, Z]$, averaged over Z, which is a summary of how D affects Y when the effect $\gamma_0(1, Z) - \gamma_0(0, Z)$ depends on Z.

There are many other examples of interesting parameters that depend linearly on a regression.

Example 3: For $\gamma_0(X) = E[Y|X]$ and known CDF's F_1 and F_2,

$$\theta_0 = \int \gamma_0(x) d\mu(x),\ \mu(x) = F_1(x) - F_0(x).$$

This is meant to be interpreted as a "policy effect" of changing the CDF of the regressors X from F_0 to F_1 while $\gamma_0(X)$ remains $E[Y|X]$; a main example of Chernozhukov, Newey, and Singh (2022).

In both Examples 2 and 3 the parameter of interest is a linear function of the conditional mean $E[Y|X]$.

There are many other interesting examples of such linear functions of a conditional mean.

A general formula for such a parameter is

$$\theta_0 = E[m(W, \gamma_0)],$$

where $W = (Y, X)$ and $m(W, \gamma)$ is a function of a data observation and a possible regression function γ, with $\gamma_0(X) = E[Y|X]$.

For Example 2, $X = (D, Z)$ and $m(W, \gamma) = \gamma(1, Z) - \gamma(0, Z)$ and for Example 3 $m(W, \gamma) = \int \gamma(x) d\mu(x)$.

This framework is useful because there is a general approach to DML for many such parameters.

$$\theta_0 = E[m(W, \gamma_0)].$$

We can construct a Neyman orthogonal estimating equation for any such parameter under a key condition.

The key condition is that there exists $\alpha_0(X)$ such that $E[\alpha_0(X)^2] < \infty$ and

$$E[m(W, \gamma)] = E[\alpha_0(X)\gamma(X)] \text{ for all } \gamma \text{ with } E[\gamma(X)^2] < \infty. \ (KEY)$$

In Example 2

$$\alpha_0(X) = \frac{D}{\pi_0(Z)} - \frac{1 - D}{1 - \pi_0(Z)}, \ \pi_0(Z) = \mathsf{Pr}(D = 1|Z).$$

Here the key condition, that $E[m(W, \gamma)] = E[\alpha_0(X)\gamma(X)]$ for all $\gamma(X)$ with finite second moment, can be shown by iterated expectations.

In Example 3, $\alpha_0(X) = f(X)^{-1}[f_1(X) - f_0(X)]$, where $f(X)$ is pdf of X.

The key condition is

$$E[m(W,\gamma)] = E[\alpha_0(X)\gamma(X)] \text{ for all } \gamma \text{ with } E[\gamma(X)^2] < \infty. \ (KEY)$$

In functional analysis this is known as a Riesz representation of $E[m(W,\gamma)]$ as a function (or functional) of γ.

We refer to $\alpha_0(X)$ as the "Riesz representer".

Existence of $\alpha_0(X)$ satisfying this key condition is known to be necessary for existence of a $\sqrt{n}$-consistent, locally regular estimator of $\theta_{0.} = E[m(W,\gamma_0)]$.

Note that by iterated expectations, the key condition implies

$$\theta_0 = E[m(W,\gamma_0)] = E[\alpha_0(X)\gamma_0(X)] = E[\alpha_0(X)Y],$$

so parameter is an "average linear combination" of Y, with "coefficient" $\alpha_0(X)$.

So existence of a $\sqrt{n}$-consistent estimator implies θ_0 is an "average linear combination" of Y.

The key condition is

$$E[m(W,\gamma)] = E[\alpha_0(X)\gamma(X)] \text{ for all } \gamma \text{ with } E[\gamma(X)^2] < \infty. \ (KEY)$$

The Neyman orthogonal moment function is

$$\psi(W,\gamma,\alpha,\theta) = m(W,\gamma) - \theta + \alpha(X)[Y - \gamma(X)].$$

This function is Neyman orthogonal for *any* functions γ and α,

$$\begin{aligned} E[\psi(W,\gamma,\alpha,\theta_0)] &= E[m(W,\gamma)] - \theta_0 + E[\alpha(X)\{Y - \gamma(X)\}] \\ &= E[m(W,\gamma)] - E[m(W,\gamma_0)] + E[\alpha(X)\{Y - \gamma(X)\}] \\ &= E[m(W,\gamma - \gamma_0)] + E[\alpha(X)E[Y - \gamma(X)|X]] \\ &= E[\alpha_0(X)\{\gamma(X) - \gamma_0(X)\}] + E[\alpha(X)\{\gamma_0(X) - \gamma(X)\}] \\ &= -E[\{\alpha(X) - \alpha_0(X)\}\{\gamma(X) - \gamma_0(X)\}]. \end{aligned}$$

where the fourth equality follows by KEY and iterated expectations.

Thus for ALL γ and α,

$$0 = E[\psi(W,\gamma,\alpha_0,\theta_0)] = E[\psi(W,\gamma_0,\alpha,\theta_0)].$$

Thus, the moment condition is satisfied at EVERY γ when $\alpha = \alpha_0$ and at EVERY α when $\gamma = \gamma_0$; this is stronger than Neyman orthogonality and has a special name of *double robustness*.

DML WITH CROSS-FITTING

We use cross-fitting to construct the estimator, where we partition the observation indices ($i = 1, ..., n$) into L groups I_ℓ, ($\ell = 1, ..., L$).

Let $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ be constructed from observations NOT in I_ℓ.

We form sample moment function $\hat{\psi}(\theta) = \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \theta)/n$, choose $\hat{\theta}$ so that $\hat{\psi}(\hat{\theta}) = 0$, and solve for $\hat{\theta}$ to obtain

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}.$$

Estimate asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ by

$$\hat{V} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_i^2, \;\; \hat{\psi}_i = m(W_i, \hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)].$$

Under regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V), \;\; \hat{V} \xrightarrow{p} V.$$

Do not have to adjust variance estimator $\hat{V}$ for $\hat{\gamma}$ or $\hat{\alpha}$ because moment function is Neyman orthogonal.

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{ m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)] \}.$$

To construct $\hat{\theta}$ we need estimators $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$; we can take $\hat{\gamma}_\ell$ to be ML of $E[Y|X]$, e.g. Lasso, Neural Net (NN), or Random Forests (RF).

Could construct $\hat{\alpha}_\ell$ by plugging in estimators of unknown objects in $\alpha_0(X)$.

For example, for ATE we could use $\hat{\alpha}_\ell(X) = D/\hat{\pi}_\ell(Z) - (1-D)/[1-\hat{\pi}_\ell(Z)]$, where $\hat{\pi}_\ell(Z)$ is an estimator of propensity score $\pi_0(Z) = \mathsf{Pr}(D = 1|Z)$.

This $\hat{\alpha}_\ell(X)$ could be unstable for small estimated propensity score.

There is an alternative way to construct $\hat{\alpha}_\ell(X)$ that avoids this instability is automatic in the sense that it only depends on the parameter of interest, through $m(W, \gamma)$, and and does not require knowing the form of $\alpha_0(X)$.

REISZ REGRESSION: AUTOMATIC ESTIMATION OF α_0

To form DML $\hat{\theta}$ we need an estimator of $\alpha_0(X)$; recall that for any $\alpha(X) = \gamma(X)$,

$$E[m(W, \alpha)] = E[\alpha_0(X)\alpha(X)] \text{ for all } \alpha(X) \text{ with } E[\alpha(X)^2] < \infty. \ (KEY)$$

We can use this equation to get an objective function that is minimized at α_0, and so can be used to form an extremum estimator of α_0.

$$\begin{aligned} E[\{\alpha(X) - \alpha_0(X)\}^2] &= E[\alpha(X)^2 - 2\alpha(X)\alpha_0(X) + \alpha_0(X)^2] \\ &= E[\alpha(X)^2] - 2E[\alpha_0(X)\alpha(X)] + C \\ &= E[\alpha(X)^2] - 2E[m(W, \alpha)] + C \\ &= E[\alpha(X)^2 - 2m(W, \alpha)] + C, \end{aligned}$$

where $C = -E[\alpha_0(X)^2]$ is a constant that does not depend on α and the third line follows by the KEY condition.

Clearly α_0 minimizes $E[\alpha(X)^2 - 2m(W, \alpha)]$, i.e.

$$\alpha_0 = \arg\min_{\alpha} E[\alpha(X)^2 - 2m(W, \alpha)].$$

We refer to this as *Reisz regression*. See Chernozhukov, Newey, Quintas-Martinez, Syrgkanis (2021).

$$\alpha_0 = \arg\min_{\alpha} E[\alpha(X)^2 - 2m(W, \alpha)].$$

We can estimate α_0 by replacing expectation with sample average and minimizing

$$\hat{\alpha}_\ell = \arg\min_{\alpha \in \mathcal{A}} \{ \sum_{i \notin I_\ell} [\alpha(X_i)^2 - 2m(W_i, \alpha)] + \hat{P}_\ell(\alpha) \}$$

over some approximating set $\mathcal{A}$ of functions for a penalty term $\hat{P}_\ell(\alpha)$, similar to Lasso.

In Example 2,

$$\hat{\alpha}_\ell = \arg\min_{\alpha \in \mathcal{A}} \{ \sum_{i \notin I_\ell} [\alpha(D_i, Z_i)^2 - 2\{\alpha(1, Z_i) - \alpha(0, Z_i)\}] + \hat{P}_\ell(\alpha) \}.$$

Different versions of Riesz regression come from choosing the set of functions $\mathcal{A}$ to minimize over and choosing the penalty.

Choosing $\mathcal{A}$ to be linear combinations of some basis of functions and the penalty to be the sum of absolute value of coefficients gives Lasso version; see Chernozhukov, Newey, and Singh (2022, EMA).

Choosing $\mathcal{A}$ to be neural net gives neural net version; see Chernozhukov, Newey, Quintas-Martinez, Syrgkanis (2021, arxiv.)

NEURAL NETS (DEEP LEARNING)

Neural Nets is a nonlinear regression specification with many parameters that is used to estimate $E[Y|X]$.

Gives remarkably good out of sample predictions in a variety of settings.

Uses as regressors a common monotonic transformation of different linear combinations of regressors, where the regressors may themselves have that form.

To describe let x denote a possible realization of X.

General neural network, nonlinear regression takes the form

$$x \stackrel{f_1}{\longmapsto} H^{(1)} \stackrel{f_2}{\longmapsto} \ldots \stackrel{f_m}{\longmapsto} H^{(m)}$$

where each layer $H^{(\ell)} = \{H_k^{(\ell)}\}_{k=1}^{K_\ell}$ has K_ℓ "neurons" $H_k^{(\ell)}$, and the function f_ℓ maps one layer of neurons to the next as in

$$f_\ell : v \stackrel{f_1}{\longmapsto} \{H_k^{(\ell)}(v)\}_{k=1}^{K_\ell} := (1, \{\sigma(v'\alpha_{k,\ell})\}_{k=2}^{K_\ell}),$$

where each $\alpha_{k,\ell}$ is a $K_{\ell-1}$ vector of parameters and where $\sigma(u)$ is a nonlinear "activation function."

Most common and theoretically interesting $\sigma(u)$ is the RELU function

$$\sigma(u) = \mathsf{max}\{0, u\}.$$

Example of single layer neural net: Approximates $E[Y|X = x]$, where x includes a constant, by

$$\alpha_{1,2} + \sum_{k=2}^{K_1} \alpha_{k,2}\sigma(x'\alpha_{k,1}).$$

Neural net approximation uses linear combination of a nonlinear transformation of a variety of different linear combinations.

Deep neural nets have more than one level, where the x variables in the above expression are themselves linear combinations of a nonlinear transformation of linear combinations, which are linear combinations of a nonlinear transformation, etc.

One layer neural net estimator is $\hat{\gamma}(x) = \hat{\alpha}_{1,2} + \sum_{k=2}^{K_1} \hat{\alpha}_{k,2}\sigma(x'\hat{\alpha}_{k,1})$ where

$$\hat{\alpha} = \mathsf{arg} \min_{\alpha_{1,2},\alpha_{2,2},\ldots,\alpha_{K_1 1},} \sum_{i=1}^{n} [Y_i - \alpha_{1,2} - \sum_{k=2}^{K_1} \alpha_{k,2}\sigma(X_i'\alpha_{k,1})]^2/n.$$

Algorithms for doing least squares to estimate neural net often involve drawing initial values for the $\alpha's$ from some distribution and then do gradient descent for OLS objective function without iterating to convergence, repeating for different starting values, and then averaging the results across different starting values.

Not iterating to convergence is a form of regularization that is useful when there are many $\alpha's$; MATLAB and R have routines for neural nets; also can regularize with ridge penalty.

There is recent theory giving mean square convergence rates, i.e. rates at which $\|\hat{\gamma} - \gamma_0\|$ converges to zero.

Rates of convergence depend on depth and width of the network and how many derivatives of $E[Y|X = x]$ exist. Depth of network is number of layers; width is maximum number of neurons (i.e. $\max_\ell K_\ell$).

Two good references are

Schmidt-Hieber, J. (2020): "Nonparametric Regression Using Deep Neural Networks with RELU Activation Function," *The Annals of Statistics* 48, 1875–1897.

Farrell, M., T. Liang, S. Misra (2021): "Deep Neural Networks for Estimation and Inference," *Econometrica* 89, 181–213.

MONTE CARLO

Chernozhukov, Newey, Quintas-Martinez, Srykganis (2022, "RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests", ICML) report results from one Monte Carlo study comparing the automatic DML with inverse propensity score method.

Do this for neural net estimator of propensity score and for neural net Riesz regression.

The NN automatic estimator of $\alpha_0(x)$ is given by

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \sum_i \{\alpha(X_i)^2 - 2[\alpha(1, Z_i) - \alpha(0, Z_i)]\} + \hat{P}(\alpha).$$

where $\mathcal{A}$ denotes neural net for $\alpha(X)$ and $\hat{P}(\alpha)$ is a penalty.

Monte Carlo comparison is with state of the art inverting of NN propensity score estimate, no cross fitting.

Estimates include some other refinements to the DML of the ATE also.

Monte Carlo design is 1000 semi-synthetic data sets based on the Infant Health and Development Program (IHDP).

IHDP is a randomized control trial about effect of home visits and attendance at specialized clinics on future developmental and health outcomes for low birth weight, premature infants (Gross, 1993).

Data using NPCI package in R under setting "A" (Dorie, 2016).

747 observations of an outcome Y, a binary treatment T, and 25 continuous and binary confounders X.

Debiasing function estimator $\hat{\alpha}$ is same as Shi et al. (2019) except use Riesz regression instead of inverse of NN estimator of propensity score $\mathsf{Pr}(D = 1|Z)$.

Results of Monte Carlo using NN for outcome and Riesz regression.

Median Absolute Error $\pm$ std. error

Riesz Reg	Dragon Net
$.110 \pm .003$	$.146 \pm .010$

Also coverage probability of nominal 95% confidence interval was .950.

In this study NN Riesz regression is better than inverting estimator of propensity score $\mathsf{Pr}(D = 1|Z)$ when Z is high dimensional.

See Chernozhukov, Newey, Singh (2022, Econometrica) for applications.

ASYMPTOTIC THEORY

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_i^2,\ \hat{\psi}_i = m(W_i,\hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)].$$

Assumption 1: $m(w,\gamma)$ *is linear in* γ.

Assumption 2: *There is a constant* $C > 0$ *such that i)* $E[m(W,\gamma)^2] \leq CE[\gamma(X)^2]$ *for all* γ, *ii)* $\alpha_0(X)$ *is bounded, and iii)* $E[Y^2|X]$ *is bounded.*

Let $\|\gamma\|_2 = \sqrt{\int \gamma(x)^2 F_X(dx)}$.

Assumption 3: $\|\hat{\gamma} - \gamma_0\|_2 \xrightarrow{p} 0$, $\|\hat{\alpha} - \alpha_0\|_2 \xrightarrow{p} 0$.

Assumption 4: $\sqrt{n}\,\|\hat{\gamma} - \gamma_0\|_2\,\|\hat{\alpha} - \alpha_0\|_2 \xrightarrow{p} 0$.

Define $V = E[\psi(W)^2]$ for $\psi(W) = m(W,\gamma_0) - \theta_0 + \alpha_0(X)[Y - \gamma_0(X)]$.

Theorem 1: *If Assumptions 1-4 are satisfied then* $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{p} N(0,V)$ *and* $\hat{V} \xrightarrow{p} V$.

–Note that i) implies that $|E[m(W,\gamma)]| \leq \sqrt{C}\,\|\gamma\|_2$ by Cauchy-Schwartz inequality which implies the KEY condition by Riesz representation theorem.

–Theorem 1 specializes Lemma 8 of "Locally Robust Semiparametric Estimation" to linear functions of the conditional mean.

–The conditions are quite simple, only depending on mean square convergence conditions.

–Mean square convergence rate for automatic Lasso estimator of $\alpha_0(X)$ given in Chernozhukov, Newey, and Singh (2022) and for the automatic NN estimator in Chernozhukov, Newey, Quintas-Martinez, Syrgkanis (2024, arxiv).

–$\alpha_0(X)$ bounded is a strong condition; for ATE this requires the propensity score $\pi_0(Z) = \mathsf{Pr}(D = 1|Z)$ bounded away from 0 and 1; for Lasso $\hat{\gamma}$ and $\hat{\alpha}$ can relax α_0 bounded to $E[\alpha_0(X)^2] < \infty$; for ATE this is $E[\pi_0(Z)^{-1}] < \infty$ and $E[(1-\pi_0(Z))^{-1}] < \infty$; minimal conditions.

–$\sqrt{n}\,\|\hat{\gamma} - \gamma_0\|_2\,\|\hat{\alpha} - \alpha_0\|_2 \xrightarrow{p} 0$ in Assumption 4 is "rate double robustness" condition that allows a trade-off in rate of convergence of $\hat{\gamma}$ and $\hat{\alpha}$.

–Assumption 4 is stronger than needed; can get root-n consistency with only $\|\hat{\gamma} - \gamma_0\| = o_p(n^{-1/4})$ and approximate sparsity; see Bradic, Chernozhukov, Newey, and Zhu (2022, arxiv).